D-SPIN – Eine Infrastruktur für deutsche Sprachressourcen

von Christina Bankhardt

Empirisch arbeitende Linguisten und Linguistinnen führen ihre Forschung auf der Grundlage immer größer werdender Datenmengen durch. Viele dieser sprachbasierten Korpora liegen derzeit verstreut und teilweise unerschlossen bei einzelnen Wissenschaftlern oder wissenschaftlichen Einrichtungen. Diese Sprachressourcen bestehen primär aus Texten oder aufgezeichneten Gesprächen und werden meist aufwändig durch teure Investitionen erhoben. In oft jahrelanger Arbeit werden sie für linguistische Bedürfnisse aufbereitet, mit Annotationen versehen, Gespräche werden transkribiert und schließlich Tools entwickelt, mit deren Hilfe die Ressourcen im Rahmen von Forschungsprojekten oder auch für die wissenschaftliche Öffentlichkeit zugänglich und nutzbar gemacht werden. Eine allgemeine und nachhaltige (Nach-)Nutzung ist jedoch aus verschiedenen Gründen erschwert oder sogar unmöglich.

Zum einen gibt es Korpora, deren Dokumentation oft unzureichend ist, da diese Arbeit in der Forschungspraxis nicht honoriert wird. Deshalb ist zum Teil völlig unbekannt, welche Ressourcen überhaupt existieren. Nur ein Bruchteil davon ist über das Web recherchierbar bzw. zugänglich. Andere Ressourcen werden gar nicht erst online zugänglich gemacht, aus Angst, sie könnten in die falschen Hände geraten, oder weil unklar ist, welche rechtlichen Konsequenzen das haben könnte. Daher bleibt ein Großteil von aufwändig erhobenen Primär- und Sekundärdaten ungenutzt. Ähnliches gilt für Werkzeuge, die sinnvollerweise auf diese Daten anwendbar sein sollten.

Zum anderen gibt es mittlerweile eine Vielzahl von recherchierbaren linguistischen Datenbanken, die Sprachkorpora vorhalten. Aber man muss sich jeweils für den Zugriff auf die Daten separat registrieren, ggf. die zugehörige Kennung und das entsprechende Passwort merken. Die Aufbereitung und Annotationsgrade dieser Daten sind zudem sehr unterschiedlich, und eine Interoperabilität oder gar Vernetzung unter den Daten und Werkzeugen an unterschiedlichen Orten ist in der Regel nicht gegeben.

Diese unübersichtliche Vielzahl von Ressourcen und Tools soll nun im Rahmen des Infrastruktur-Projektes D-SPIN (Deutsche Sprachressourcen-Infrastruktur, <www.sfs.uni-tuebingen.de/dspin/>) lokalisiert und gebündelt werden; die Arbeit vieler Jahre und großer Investitionen soll gesichert werden. Das IDS, insbesondere die Projekte COSMAS II und „Ausbau und Pflege der Korpora geschriebener Gegenwartssprache“, kooperiert im Rahmen von D-SPIN, ein vom BMBF (Bundesministerium für Bildung und Forschung) und MWK-BW (Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg) gefördertes Projekt im Rahmen des EU-Projekts CLARIN (Common Language Resources and Technology Infrastructure Network, <www.clarin.eu/>), auf deutscher Ebene mit folgenden Partnern: MPI (Max-Planck-Institut) für Psycholinguistik mit Sitz in Nijmegen (NL), Berlin-Brandenburgische Akademie der Wissenschaften, DFKI (Deutsches Forschungszentrum für Künstliche Intelligenz) Saarbrücken und den Universitäten Tübingen, Leipzig, Frankfurt, Stuttgart und Gießen.

Ziel des Projektes ist der Aufbau einer Forschungsinfrastruktur für die Sprachwissenschaft. Zukünftig soll es mit nur einer Registrierung und damit auch nur einer Benutzerkennung möglich sein, auf sämtliche Sprachressourcen der Projektpartner nach einer Autorisierung bzw. elektronischen Unterzeichnung der notwendigen Lizenzverträge zugreifen zu können. Dafür werden Standards für den Umgang mit Sprachressourcen entwickelt. Die einzelnen Ressourcen und Werkzeuge werden deshalb im Rahmen des Projektes erschlossen, harmonisiert und nachhaltig für die Wissenschaft verfügbar gemacht. Für das IDS bedeutet dies eine Anpassung seiner Ressourcen und Werkzeuge, insbesondere z. B. die Aufbereitung seiner Sprachkorpora sowie die Erweiterung der Recherche- und Analysesoftware COSMAS II.

Darüber hinaus werden am IDS die rechtlichen und ethischen Rahmenbedingungen bei der Verwendung von Sprachressourcen geklärt. In Bezug auf die rechtlichen Fragestellungen geht es zunächst um eine Bestandsaufnahme der offenen Fragen im Hinblick auf urheber- und lizenzrechtliche Probleme der Textkorpora. Die Texte wurden teilweise von mehreren Autoren verfasst, Verleger haben meist die ausschließlichen Nutzungsrechte dieser Texte inne. Eine weitere

rechtliche Ebene in diesem Konglomerat bilden die Annotationen, die im Rahmen der linguistischen Aufbereitung von Wissenschaftlern bzw. „Taggern“ und anderen, insbesondere kommerziellen Werkzeugen hinzugefügt werden.

Außerdem können auch bei den Korpora der gesprochenen Sprache eine Vielzahl von Individuen an einer Sprachressource geistiges Eigentum erworben haben. Hinzu kommen speziell bei Gesprächsdaten persönlichkeits- und datenschutzrechtliche Aspekte, die bei einer öffentlichen Zurverfügungstellung der Daten beachtet werden müssen. Hier liegt auch ein Schwerpunkt der ethischen Gesichtspunkte. Inwieweit ist es bspw. moralisch vertretbar, ein Therapiegespräch zwischen einem Arzt und seinem Patienten für die Gesprächsforschung über das Internet bereitzustellen?

Vor dem Hintergrund dieses komplexen Terrains aus rechtlichen und ethischen Fragestellungen geht es bei D-SPIN um die Entwicklung von Best-Practice-Richtlinien für den Umgang mit Sprachressourcen. Es sollen Musterverträge und Lizenzmodelle, auf europäischer Ebene im Rahmen des CLARIN-Projekts koordiniert, erarbeitet sowie ethisch und rechtlich gebotene Einschränkungen der Zugriffsrechte auf die einzelnen Ressourcen untersucht werden.

In Zukunft sollen dadurch aufwändige Erhebungen erleichtert bzw. überflüssig gemacht und die Vergleichbarkeit von Untersuchungen und wissenschaftlichen Ergebnissen gesteigert werden, und zwar aufgrund einer größeren Transparenz der Forschungsergebnisse, da sie durch zugängliche Primär- und Sekundärdaten besser verifiziert bzw. falsifiziert werden können.

Seit Anfang 2009 wird dieser IDS-Anteil von D-SPIN von dem neuen Projekt „Aufbau eines Zentrums ‚Digitale Forschungsressourcen für die germanistische Sprachwissenschaft‘“ mit anderen IDS-Aktivitäten im Umfeld von Forschungsinfrastrukturen, eHumanities und der (Nach-)Nutzbarmachung von Forschungsressourcen koordiniert.

Veranstaltungshinweis:

Am 15. und 16. Mai 2009 wird am IDS in Kooperation mit den Kollegen aus Tübingen ein Workshop im Rahmen des D-SPIN-Projektes stattfinden.

Die Autorin ist wissenschaftliche Mitarbeiterin am Institut für Deutsche Sprache in Mannheim.